**Working Paper**

Statistical Matching of the EWCS and EQLS

**Author:** Mathijn Wilkens

Contents

[Introduction 1](#_Toc33184450)

[1 – Statistical matching in theory 2](#_Toc33184451)

[Definition 2](#_Toc33184452)

[Assumptions 2](#_Toc33184453)

[Approach 3](#_Toc33184454)

[2: Choice of the target variables 6](#_Toc33184455)

[3: Comparing EQLS and EWCS 7](#_Toc33184456)

[Coverage and universe 7](#_Toc33184457)

[Fieldwork 8](#_Toc33184458)

[Survey methodology 10](#_Toc33184459)

[Identification of common variables 10](#_Toc33184460)

[5: Selection of matching variables 13](#_Toc33184461)

[Accuracy 13](#_Toc33184462)

[Variable importance 14](#_Toc33184463)

[Conclusion 16](#_Toc33184464)

[6: Statistical Matching 17](#_Toc33184465)

[7: Evaluation 20](#_Toc33184466)

[Preserving distributions 20](#_Toc33184467)

[Association between matched variables 20](#_Toc33184468)

[Conclusions 23](#_Toc33184469)

[References 24](#_Toc33184470)

[Variables 25](#_Toc33184471)

[EWCS variables (Y) 25](#_Toc33184472)

[EQLS variables (Z) 25](#_Toc33184473)

# Introduction

Eurofound’s European Working Conditions Survey (EWCS) and European Quality of Life Survey (EQLS) are carried out separately, but share many of the same or similar questions. In addition, some questions are only observed in the EWCS and others only in the EQLS. Combining the overlapping samples by statistical matching allows us to jointly observe these variables and thereby enrich both surveys with a range of variables free of cost.

However, the respondents in the EWCS sample are not the same as in the EQLS sample, so statistical matching is only feasible to the extent we are willing to accept the assumption that respondents are the same person, given a set of characteristics of those persons. This working paper explores the potential of statistically matching the EWCS and EQLS, assesses its feasibility and outlines the procedure and results.

# 1 – Statistical matching in theory

This section summarises the concept of statistical matching and discusses the main assumptions and approaches. This section is based on D’Orazio et al (2006), Eurostat (2013) and D’Orazio (2015).

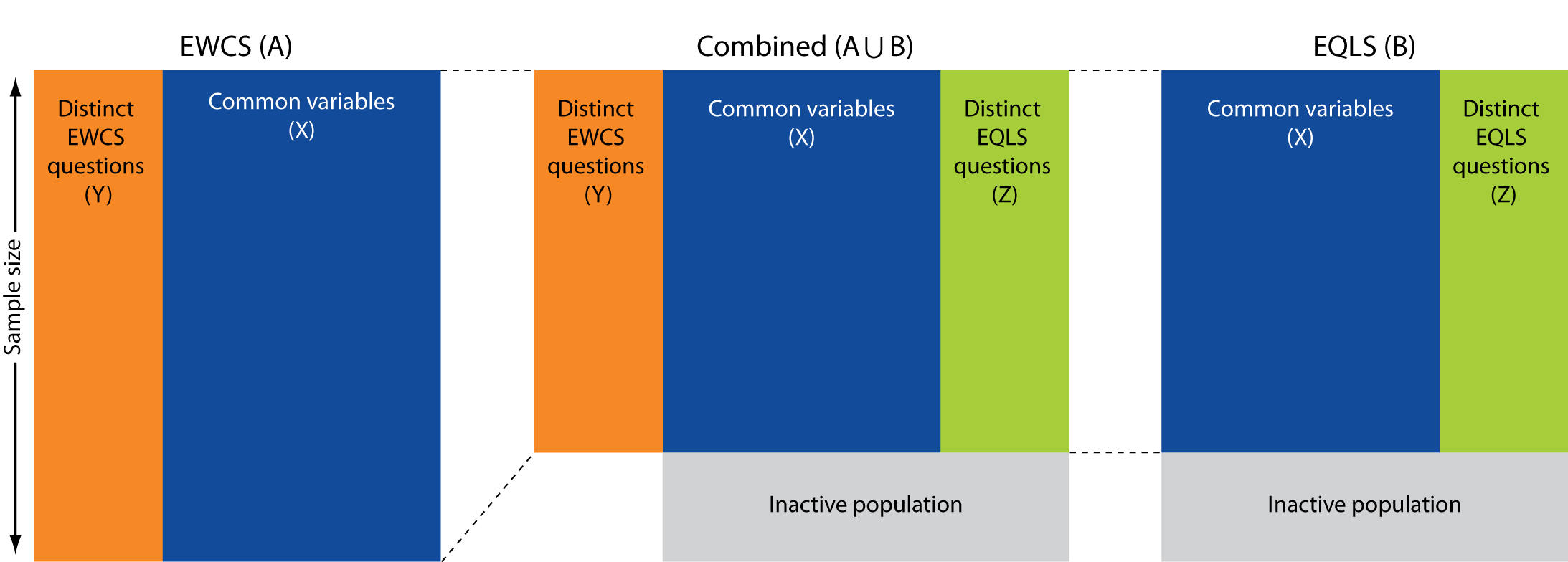
## Definition

Statistical matching is as a data integration procedure that aims to integrate two or more datasets characterised by the fact that:

* The different data sets contain information on a set of common variables and variables that are not jointly observed
* The units observed in the data sets are different (disjoint sets of units).

Both conditions apply to the EWCS and EQLS. The surveys share a relatively large set of common variables, hence the potential of statistical matching. Also, both surveys have variables that are not present in the other survey and statistical matching would allow us to create a synthetic sample of respondents that have values for all variables.

Figure 1: Statistical matching of the EWCS and EQLS.



## Assumptions

A core assumption of the statistical matching framework is that the two surveys that are going to be matched together are two independent samples from the same distribution. In other words: we assume that both surveys are random subsamples from the same survey. We need this assumption for doing any statistical inference on the synthetic dataset. In reality, this is a strong assumption. For example, the fieldwork of the 6th EWCS was conducted in 2015, while the fieldwork of the 4th EQLS was conducted in late 2016 and early 2017. The question is to what extent it is reasonable to assume that the 6th EWCS and the 4th EQLS can be regarded as random samples from the same distribution. This is not just affected by the timing of the fieldwork: for example sampling or fieldwork practices may differ between the surveys or non-response bias may differ. In addition, the population is only partially overlapping, because the EWCS only covers people in paid employment, while the EQLS is a general population survey. The statistical matching therefore only applies to the working respondents.

Another crucial assumption is the Conditional Independence Assumption (CIA). If we have a set of variables X, Y and Z, and (X,Y) is observed in dataset A and (X,Z) in dataset B, we could create a synthetic dataset A ∪ B with variables (X,Y,Z) by statistical matching. However, we need to assume that Y is independent from Z given X, because otherwise a model that matches both datasets would not be identifiable. We match variable Z from dataset B to dataset A (or vice versa). Given a certain value of X, we assume that the probability of Z does not affect the probability of Y.

For example, variables on the use of child care services in the EQLS may not be independent from the availability of flexible work arrangements as captured in the EWCS. Perhaps those who use child care services are more likely to work in organisations that offer flexible working time arrangements. This would violate the independence assumption, but perhaps not the CIA if we assume that a range of variables X, for example age, gender, household composition, profession, working hours and work-life balance may explain the probability of using of child care services as well as the probability of working in an organisation with flexible working hours. Conditional on these variables, the use of child care services and the availability of flexible work arrangements can then be considered independent. This does not mean that all respondents that use child care have the same availability of flexible work arrangements. Any residual variance in flexible working arrangements does simply not depend on the use of child care, but on other factors independent from the use of child care. For example, there may not be any child care facilities available in the area where the respondent lives, although there is variation in the flexible work arrangements that organisations offer.

## Approach

Statistical matching generally requires five steps:

1. Choice of the target variables Y and Z observed distinctly in the two surveys
2. Identification of common variables X shared by A and B.
3. Selection of matching variables X.
4. Statistical matching of A and B, creating a combined dataset with X,Y and Z.
5. Evaluation of the matching.

The choice of the target variables Y and Z depends on the research question and the availability of data. The reason for statistically matching datasets A and B is observing the joint distribution of Y and Z. The joint distributions (X,Y) and (X,Z) can already be observed in datasets A and B respectively. The choice of the target variables Y and Z matter for the subsequent steps in the statistical matching procedure, in particular the choice of the matching variables. It is unlikely that any combination of matching variables will be adequate to match all distinct variables in dataset A to all distinct variables in dataset B.

After the choice of target variables has been made, the second step consists of the identification of common variables X. Ideally, definitions and classifications are identical in both datasets. In practice, this rarely holds true completely. Comparing confidence intervals for example can assess the extent to which we can assume the variables in both surveys measure the same concept. Reconciliation steps actions can be undertaken to harmonise the common variables. This may include transforming the units (e.g. currency conversion), transforming categories (e.g. merging items on a likert scale), recoding combinations of variables into new variables or handling missing data.

Not all identified common variables X need to be used for matching. After identifying all possible candidates in step 2, we can proceed with selection the relevant matching variables in step 3. This depends on step 1 – the target variables Y and Z. From a computational and inferential point of view, a smaller number of X variables is preferred. However, more X variables may make the CIA assumption more plausible. The right balance between these two ends is the selection of variables X such that the common variables other than the matching variables X are independent of Y given X and independent of Z given X. This can be tested in a multivariate setting by regression Y and Z on the common variables. Those common variables that correlate significantly are selected as matching variables. More sophisticated methods are stepwise regression, hierarchical cluster analysis in the case of multicollinearity, or classification and regression trees in the case of non-linearity.

Once we have established the appropriate matching variables in step 3, we can proceed with the actual statistical matching. The most common and flexible approach for statistical matching on micro-level is hot deck imputation. The general idea is that missing values from a recipient file A is replaced by values from a donor file B. The donor file is usually the biggest file, because otherwise some records in the donor file may be imputed more than once in the recipient file. This implies that the EWCS is the donor file and the EQLS the recipient file because although the sample sizes are comparable, the overlapping sample is limited to the working population (figure 1).

There are three hot deck methods

1. Random hot deck: random draws from the donor file within a suitable subset of units. E.g. random draw from all men with a certain age in a certain country, etc.
2. Rank hot deck: if one of the matching variables is ordinal, units are matched based on rank. E.g. the highest income in set A is matched with highest income in set B.
3. Distance hot deck: records are matched according to the smallest distance, e.g. the income in the donor record that comes closest to the recipient record in terms of the difference in Euros. Distance hot deck can be applied constrained and unconstrained. Constrained is when a donor record can only be used once; the advantage is that the marginal distribution of the donor file will be maintained. The disadvantage is that large distances may need to be accepted. The alternative is the unconstrained distance hot deck, where donors can be used multiple times.

Given finite sample sizes, there may not be donors that adequately resemble the recipients. The additional variance and distortion due to the non-coincidence between the donor and the recipient is called matching noise.

After the synthetic data is created, the final step involves the evaluation of the statistical matching procedure. Four levels of validity can be assessed:

1. The marginal and joint distributions of variables in the donor sample are preserved in the statistical matching file.
2. The correlation structure and higher moments of the variables are preserved after statistical matching.
3. The true joint distribution of all variables is reflected in the statistical matching file.
4. The true but unknown values of the Z variable of the recipient units are reproduced.

Level 1 can be assessed by comparing marginal and joint distributions in the donor and combined datasets. Level 2 can be assessed similarly. The extent to which level 3 is satisfied is often unknown because the objective of statistical matching is usually obtaining the joint distribution of the variables distinct to separate surveys.

The use of auxiliary information can be useful for assessing whether the joint distribution reflects the true joint distribution. Matching noise may cause these to differ from each other. A source of auxiliary information could be a different dataset C containing (X,Y,Z) or (Y,Z). The information on the joint distribution of these variables can be used to assess the statistical matching of datasets A and B or identify a model that allows the matching of dataset A with (X,Y) and dataset B with (X,Z). The caveat here is the assumption that, in addition to dataset A and B, also dataset C is a subsample from the same distribution. This may not be true, i.e. when dataset C is a survey from a different year or with a different survey methodology.

Alternatively, uncertainty analysis can shed light on the uncertainty due to the matching process itself in the final estimates. There may be a range of solutions for a matching model that estimates f(X,Y,Z) and this uncertainty can be revealed by calculating confidence intervals around the estimates of joint distributions. The greater the explanatory power of the matching variables, the less uncertainty remains for creating the combined dataset. This emphasises that the selection of matching variables is crucial in statistical matching.

Finally, level 4 is only of importance when we are interested in more than the joint distribution of (X,Y,Z), i.e. the values of Z of the recipient units. This is not the case when matching EWCS and EQLS.

# 2: Choice of the target variables

The choice of the variables from each data source that will be combined into one synthetic datafile depends on the research question. In this case, statistical matching is performed in the context of a research project on work-life balance. The general question is to what extent work-life balance and work-life reconciliation is associated with characteristics of the job and care responsibilities. Distinct variables from the EWCS include questions on atypical working hours, regularity of working hours, working time arrangements and job intensity. Distinct variables in the EQLS include child care and long term care responsibilities – formal and informal – and questions on the availability of services. Both surveys contain questions on work-life balance and reconciliation of work and private life. A second objective is to explore how work and care are associated with each other. See section ‘variables’ for an overview of the variables of interest.

# 3: Comparing EQLS and EWCS

The previous section mentions that statistical matching assumes that the datasets are drawn from the same distribution. For this assumption to be reasonable in practice, this implies that the datasets need to be similar in terms of universe, variable definitions and survey methodology in general. This section outlines the difference between the EQLS and the EWCS and discusses to what extent the surveys can be considered as independent samples from the same distribution.

## Coverage and universe

The 4th EQLS was conducted in the EU28 and in Albania, FYROM, Montenegro, Serbia and Turkey. The same countries were surveyed for the EWCS with the addition of Norway and Switzerland.

Within each country, the EWCS covers persons aged 15 or over, except in Bulgaria, Spain and the UK where the age was 16 or over. However, the EQLS only covers those aged 18 or over. Both surveys cover only people whose usual place of residence is within the territory covered by the survey.

The EQLS covers working and non-working people. The EWCS, however, only covers people in employment. The EQLS measures employment by asking all respondents ‘which of these categories describes your current situation the best?’ (Q2c) and ‘these categories’ refer to the categories on a showcard (table 1). Only respondents who answer categories 1, 2, 3 or 4 are considered to be employed for the remainder of the survey.

In the EWCS, those in employment are those who did at least one hour of work for pay or profit during the week preceding the interview from Monday to Sunday. This is established in the screening phase of the interview to check whether the respondent is eligible. In the actual interview, the respondent is then also asked to describe ‘which of these categories describes your current situation the best?’ The categories differ slightly from the EQLS (table 1). Regardless of the answer, all respondents are considered to be in employment because this has already been established with the screener.

Table 1: Employment in the household grids of EQLS and EWCS

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **EQLS (Q2c)** | **#** | **EWCS (Q2c)** |
| 1 | at work as employee or employer/self-employed | 1 | at work as employee or employer/ self-employed \*/relative assisting on family farm or business \*\* |
| 2 | employed, on childcare leave | 2 | unemployed |
| 3 | employed, on other special leave (e.g. sickness; not holiday) | 3 | unable to work due to long-term illness or disability |
| 4 | in receipt of retirement pension and at work as employee or employer/self-employed | 4 | at work and on child-care leave or other leave |
| 5 | at work as relative assisting on family business or farm\* | 5 | retired |
| 6 | unemployed less than 12 months | 6 | full time homemaker/ responsible for ordinary shopping and looking after the home |
| 7 | unemployed 12 months or more | 7 | in full time education (at school, university, etc.) / student |
| 8 | unable to work due to long-term illness or disability | 8 | other (e.g. military duty )\*\*\* |
| 9 | retired |  |  |
| 10 | Full-time homemaker / fulfilling domestic tasks |  |  |
| 11 | in education (at school, university, etc.) / student |  |  |
| 12 | other (NOT ASKED/NOT ON CARD) |  |  |

The difference between EQLS and EWCS with respect to the measurement of employment is that the EQLS may have respondents in categories 5 and further that are employed according to the EWCS definition. In the EWCS there are respondents who are considered to be in employment, but would not be considered as such in the EQLS. The exception is ‘retired’, because the EQLS has a specific category for those who are in receipt of a retirement pension but whom are also in employment.

Because the EQLS definition is stricter, the EQLS definition of employment will be used for the purposes of statistical matching.

The implications of the differences in coverage and universe are:

* Statistical matching of the EWCS and EQLS will be limited to the EU28 as the project in which the synthetic data will be used is limited to the EU28.
* EWCS respondents younger than 18 will be excluded.
* Only employed respondents can be matched and those who answer 1-4 to Q2c of the EQLS will be matched to those who answer 1,4 or 5.

The EQLS has 30890 respondents in the EU28, 15519 of which are employed. Following the EQLS definition of employment, the EWCS has 34277 employed respondents.

## Fieldwork

The 6th EWCS was conducted between February and October 2015 for the EU28. The EQLS was conducted from September 2016 to March 2017 for the EU28. In addition to covering a different year and a different period within the year, fieldwork length also differs between countries in both surveys (figure 2). These differences may affect the comparability of the surveys, although it is unclear to what extent.

The response rate of the EQLS is 37% for 33 countries in total. For the EWCS, this is 43% for 35 countries. Although the overall response rate between the surveys is similar, there are marked differences between countries (figure 3). The largest differences can be found in Greece, Italy, Germany, Lithuania and Poland where the EWCS response rate is much larger than the EQLS response rate. In Spain, the reverse holds.

Differences in response rates do not necessarily affect the comparability of the surveys unless there is a difference in non-response bias. For both surveys, non-response bias is unknown, but if we assume larger differences in response rates reflect larger differences in non-response bias, we may find larger differences in distributions in the countries that have large differences in response rates.

Figure 2: Fieldwork intensity of the EQLS and EWCS.

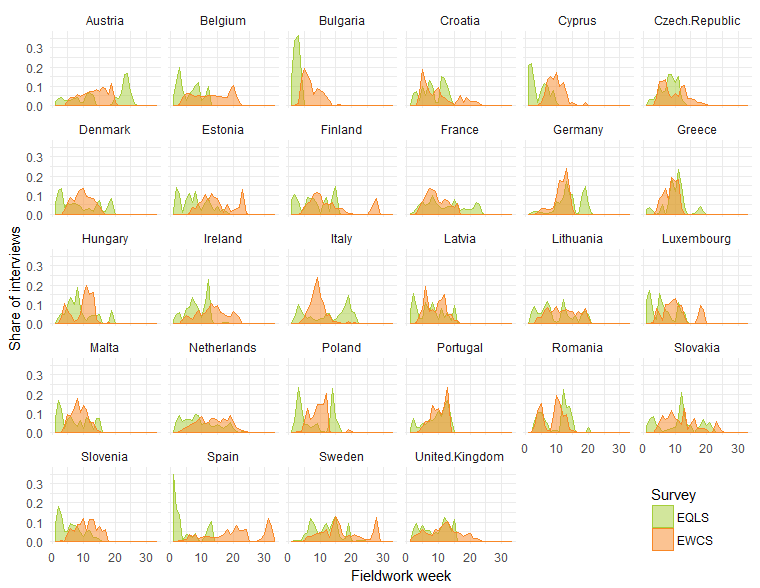
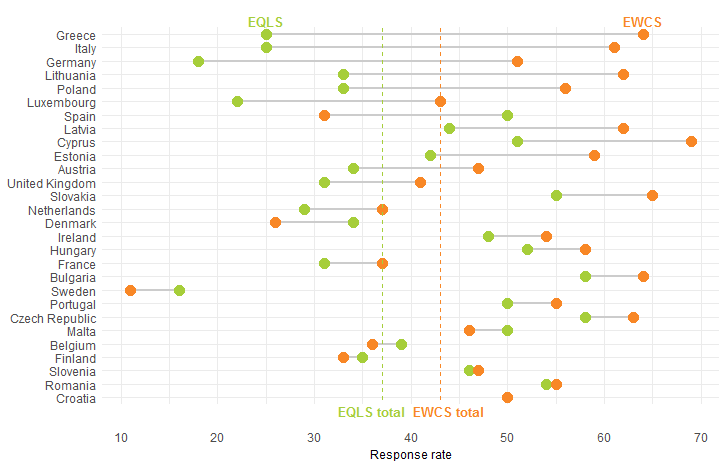


Figure 3: Response rates EQLS and EWCS



## Survey methodology

Both surveys apply random probability sampling and because the surveys are commissioned by the same organisation (Eurofound), survey methodology is largely the same. Because it is not straightforward to determine the effects of any differences in survey design on the comparability of the survey, the differences in design will not be outlined in detail. More details can be found in the EQLS and EWCS technical reports which are available on the Eurofound website.

Also, both surveys are complex surveys in the sense that they both have multiple sampling stages. Both survey sample individuals within households, within primary sampling units (PSUs), within strata, within countries. Both surveys include post-stratification weights that correct for differences in the distribution of a certain set of variables relative to reference statistics. These complex survey designs need to be taken into account by applying the weights and by correcting for clustering when estimating variance.

## Identification of common variables

The EWCS and EQLS have roughly 14 identical items in the questionnaire. In addition, 18 items can be considered similar, meaning that they are common in definition, but not in measurement. Wording may be slightly different or answering categories may be different. In some cases, recoding of categories may improve the comparability.

To assess the similarity of the variables in both surveys, we can calculate Hellinger’s distance, which is a measure of dissimilarity of the distributions. A rule of thumb for Hellingers distance is that we can consider distributions to be equal if the value is below 0.05. Hellinger’s distance is based on frequency tables which for the purposes of this exercise have been weighted. Annex ‘Identification of common variables’ presents an overview of the distributions of each variable as well as Hellinger’s distance.

Sample size differs between the survey and Hellinger’s distance does not take this into account. For testing that we would need statistical tests, but this is complicated given that we need to take into account the variance due to clustering. One option is to divide the Chi-square test by the generalised design effect of both surveys. The ‘Delta H0’ shows the value of the generalised design effect that would determine the acceptance of the null hypothesis (equality of distributions) in the case of alpha=0.05 (df=J-1). In other words: the minimum design effect we would have to assume in order to conclude the distributions are the same.

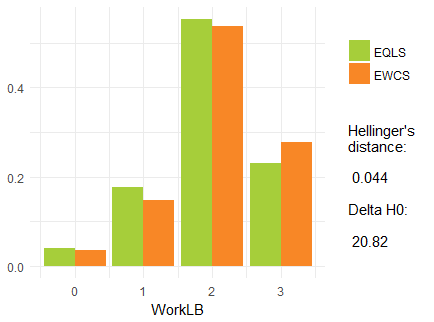
The full overview is available in the appendix, four variables are shown below as examples. Figure 4 shows the distribution of gender in the EWCS and EQLS and as expected these can be considered very similar. Hellinger’s distance is 0.01, which is below the threshold of 0.05.

|  |  |
| --- | --- |
| Figure 4: Gender in the EQLS and EWCS | Figure 5: Working hours in the EQLS and EWCS |

The question on the number of working hours in the EWCS is phrased: How many hours do you usually work per week in your main paid job? In the EQLS this is slightly different: How many hours do you normally work per week in your main job, including any paid or unpaid overtime? Figure 5 shows that the distribution is roughly the same, but not exactly. There is a higher concentration of respondents that work more than 40 hours, which may be the result of the explicit mentioning of paid or unpaid overtime in the EQLS. Mean of working hours in the EQLS is also 1.5 hours higher than in the EWCS. Hellinger’s distance is just above the threshold.

The question on work-life balance is important for the purpose of the statistical matching. In both surveys, the question is: In general, how do your working hours fit in with your family or social commitments outside work? However, in the EWCS the scale is from ‘Not at all well’, ‘Not very well’, ‘Well’, and ‘Very well’, while in the EQLS the middle two categories are ‘Rather not well’ and ‘Rather well’. Nevertheless, the distributions in both surveys are still very similar (Figure 6)

Figure 6: Work-life balance in the EQLS and EWCS



Annex ‘Identification of common variables’ shows a full overview of all the variables that the EWCS and EQLS have in common. Table 2 summarises the findings. In the assessment, not only distributions have been compared, but also similarity in the degree of association to each other by looking at similarity in correlations for continuous variables, similarity in chi square tests for categorical variables and similarity in regression coefficients of regressions of continuous on categorical variables. The final table in Table 2 concludes.

Table 2: Summary assessment similarity of common variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Question** | **Distributions** | **Correlations** | **Conclusion** |
| Gender | Same | Same | Same | **🗸** |
| Age | Same | Same | Same | **🗸** |
| Employment status | Same | Same | Same | **🗸** |
| Working time | Different | Similar | Same | **🗸** |
| Preferred hours | Same | Similar | Different | **🗴** |
| ISCO | Same | Similar | Same | **🗸** |
| Partner | Same | Same | Same | **🗸** |
| Children | Same | Same | Same | **🗸** |
| Household size | Same | Same | Same | **🗸** |
| Working hours of the partner | Same | Same | Similar | **🗸** |
| Health | Same | Same | Same | **🗸** |
| WHO5 | Same | Similar | Similar | **🗸** |
| Commuting time | Same | Same | Same | **🗸** |
| Education | Same | Same | Same | **🗸** |
| Making ends meet | Same | Same | Same | **🗸** |
| Migrant | Same | Same | Similar | **🗸** |
| Work-life balance | Different | Same | Similar | **🗸** |
| Too tired | Different | Different | Different | **🗴** |
| Family responsibilites | Different | Different | Different | **🗴** |
| Cant concentrate | Different | Different | Different | **🗴** |
| Might lose job | Different | Different | Similar | **🗴** |
| Find new job | Different | Different | Different | **🗴** |
| Care for children | Different | Different | Similar | **🗴** |
| Care for relatives | Different | Same | Similar | **🗸** |

Out of 24 variables, 17 can be considered as variables common to the EWCS and EQLS. Important to mention is that the items on work-life reconciliation (‘I have felt too tired after work to do some of the household jobs’, ‘I found that my job prevented me from giving the time I wanted to my family’ and ‘I found it too difficult to concentrate on the job because of family responsibilities’) cannot be considered similar in the EQLS and EWCS. The wording of the questions is different, the distributions are different and also the correlations with other variables show different patterns.

# 5: Selection of matching variables

The previous step (identification of common variables) resulted in a list of potential variables X that could be used for matching. The aim of this step is to assess which of those variables X are relevant for variables that are not common to both surveys, variables Y and Z (Figure 1). When considering which matching variables to choose, two conditions should be satisfied:

* The matching variables should be good predictors of the Y variables; the variables unique to the EWCS that will be matched the EQLS. Common variables that have no explanatory power whatsoever will add no value over a completely random matching procedure
* The matching variables should satisfy the conditional independence assumption. In other words, association between Y and Z controlled for X should be zero. This cannot be tested, but becomes more plausible the more the common variables are predictive of Y and Z.

Causality is not relevant for statistical matching. The point is to use matching variables that can predict well the variables that will be matched, based on a causal relationship or not.

Hot deck matching procedures are non-parametric and allow for non-linear associations with the matching variables and Y and Z. Therefore, the method to assess which common variables are good predictors of Y and Z should be flexible. Methods that have been used before are simple correlations or similar measures of associations as well as (stepwise) regressions. The drawback of using a parametric method such as regression is that non-linear parameters will have to be specified beforehand.

A more flexible, non-parametric and powerful method is random forests for classification and regression trees (Breiman, 2001). The idea is to predict the variables of interest X and Y with the common variables Z. Random forest randomly creates classification and regression trees of different combinations of common variables Z and combines the information from these trees to predict Y and Z. For selecting the matching variables, the purpose is not to be able to predict Y and Z perfectly. The goal is to assess which common variable Z have the most predictive power. To assess this, we can look at variable importance as implemented in the R-package randomForest.

In the figures below, variable numbers or names are shown. See section ‘variables’ at the end of this document for the questions corresponding to the variables.

Further details can be found in the annex ‘Selection of matching variables’.

## Accuracy

The accuracy of random forest models can be assessed by out-of-bag (OOB) error rates for categorical variables and R2 for continuous variables.

The OOB error rate shows the share of observations with a predicted value different from the observed one. The lower the rate, the more accurate the model. Lower OOB error rates can be expected for variables with fewer categories, i.e. for a variable with two categories there is a 0.5 probability of a wrong prediction by simply guessing, while this is higher for variables with more categories. The R2 shows the share of explained variation in the dependent variable.

Normally, models would be trained on a training data set and tested on a separate dataset. The argument is to avoid overfitting, i.e. to avoid a model that fits one particular data set very well, but has a weak performance in other data sets. However, because statistical matching is aimed at two datasets precisely, there is no need to train and test in separate datasets.

Figure 7 shows the accuracy of the random forest models for each variable. The common variables can predict the use of childcare, the type of childcare (Q78) and effects of costs (Q82) well. For the use of services we get reasonable predictions, e.g. for Q68a roughly 75% of the answers are predicted correctly and this is 50% for Q68b. Because these questions have three categories guessing would give us 33% accuracy.

For the continuous variables we explain roughly 17% of the variance. Some of the variables are based on sub-selections so will have a smaller sample size. The reasonably high predictive power of the common variables for most of the EQLS variables makes the CIA more plausible.

|  |  |
| --- | --- |
| Figure 7: Accuracy of random forest for EQLS variables (1-OOB and R2) | Figure 8: Accuracy of random forest for EWCS variables (1-OOB and R2) |

## Variable importance

From the random forest models, the variable importance can be extracted to show which of the common variables are most predictive of the variables of interest Y and Z. The section shows the predictive power of the combination of the variables included in the models, and below is shown what each variable contributes to this. This is measured by the mean decrease in accuracy, which measures how much more faulty predictions we would get if we would remove a particular variable from the model. The mean decrease in gini is a similar measure, but is a measurement of how important the variables are for splitting the trees. For both statistics, a higher value indicates a higher variable importance. Those variables that consistently show high variable importance should be candidates for matching variables.

For each Y and Z variable, a separate random forest model has been estimated and for each model, there are two measures of variable importance for each X variable in that model. If we take the mean of the variable importance measures (mean decrease in accuracy and the mean decrease of the gini index) over the models, we can summarise the variable importance.

Figure 9 shows the importance of the common variables X for the EWCS variables Y. Country, ISCO and working hours are the most important explanatory variables for the EWCS variables. Other variables are of less importance, and the mean decrease accuracy shows some differences with mean decrease gini, but employment status, age, mental wellbeing (WHO5), commuting time and work-life balance also seem important.

Country is also a very important predictor of the EQLS variables (Figure 10), followed up by age, making ends meet, ISCO, the number of working hours and mental wellbeing (WHO5). Generally, the differences in variable importance for the EQLS are not as outspoken as for the EWCS. Certain questions were only asked to respondents with children, explaining why having children is not an important explanatory variable.

Figure 9: EWCS variable importance in random forest. Mean of the mean decrease in accuracy and the mean of the mean decrease in gini.

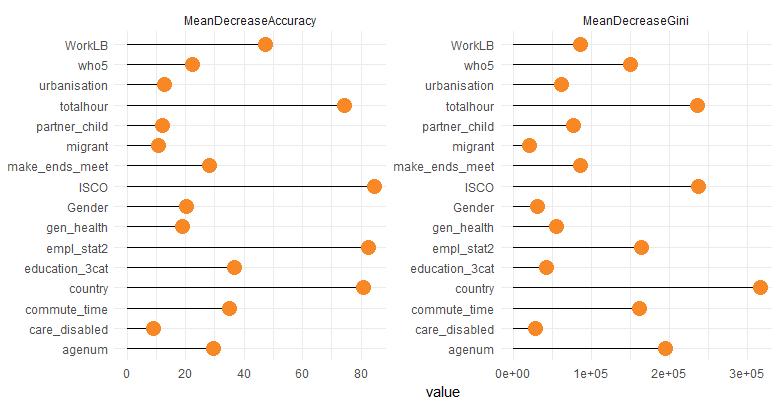
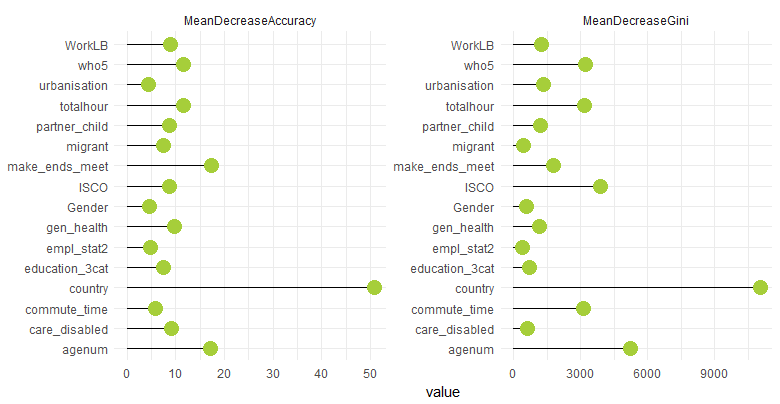


Figure 10: EQLS variable importance in random forest. Mean of the mean decrease in accuracy and the mean of the mean decrease in gini.



## Conclusion

The common variables X seem to predict Y and Z reasonably well. It is unclear how well the common variable should predict the Y and Z variables, but to satisfy the CIA assumption we need to have matching variables that explain all the association between Y and Z.

Based on the variable importance, the following order of importance will be considered for matching the EWCS variables:

1. Country, ISCO and working hours
2. Employment status, WHO5, and Work-life balance
3. Commuting time, make ends meet

# 6: Statistical Matching

Following the approach described in section 1, statistical matching of the EQLS and EWCS can be performed on the basis of the common variables mentioned in section 4. There are trade-offs for adding more variables as matching variables and solving the matching model will become computationally more difficult to more matching variables are added. Therefore, the conclusions of section 5 will be taken into consideration when choosing appropriate matching variables. This section describes the statistical matching approaches taken and section 7 evaluates the different approaches.

R-package ‘StatMatch’ is used for implementation of the statistical matching procedure. The random hot deck procedure does not allow for constraining the procedure to match a donor only once to a recipient while this is possible for rank and distance hot-deck methods. Distance hot-deck procedures allow for using more than 1 matching variable and also allow for including continuous as well as categorical variables so distance hot deck will be applied.

The distance hot deck procedure distinguishes between matching variables and donation classes. Matching variables are variables that will be used to calculate distances between the values of recipients and donors. The donor with the smallest distance in terms of the matching variables will be matched to the recipient. Note that this does not mean the donor will have exactly the same value as the recipient, it is the closest value possible based on the distance calculated in conjunction with other matching variables. Donation classes are variables that define within which classes distances will be calculated. For example, if a donation class consists of country A and B, recipients from country A will only be matched to donors from country A and to a donor that has the closest distance within country A in terms of the matching variables.

Given the importance of country as an explanatory variable, and as an element of the survey design, country is considered a donation class and matching will take place within each country. This means that respondents from a certain country in the EQLS can only be matched to respondents from the same country in the EWCS. In addition, having children is also added as a donation class because certain questions in the EQLS are addressed only to people with children. Finally, work-life balance is added as donation class because one aim is to do a regression with work-life balance as a dependent variable on the joint datafile.

The choice between what to use as a donation class or as a matching variable matters. Using variables as donation classes will ensure they have exactly the same value in the donation dataset for those variables. This is not the case for matching variables. In addition, the more donation classes we add the more likely there will not be at least one donor in the same classes and we cannot perform statistical matching. Also, even if there are enough donors, the distances of the matching variables are likely to grow. This may be acceptable, depending on how important those variables are for the CIA.

Several matching models have been implemented. For more details, see annex ‘statistical matching’. There are four model specifications and all of them use country, work-life balance and having a child in the household as donation classes:

1. Full: this model uses all the common variables as matching variables, including those which were not particularly relevant for the variables to be matched.
2. Important: this uses all the important common variables, based on the variable importance determined in the last step. ISCO, work hours, employment status, age, who5, commuting time and making ends meet.
3. Half: this is a reduced version of model 2, excluding commuting time and making ends meet.
4. Minimal: This only includes ISCO and the workhours, so only the most crucial matching variables.
5. Random: a model that only uses country as a donation class, but does not have any matching variables.

These model specifications reflect different trade-offs.

The more matching variables we add to the model, the more difficult it will be to find close donors in terms of all these variables and the distributions of EQLS recipient and EWCS donor in terms of these variables will be more different. Also, the more matching variables we add, the more we move away from a random sample and the more difficult it will be to reproduce the marginal and joint distributions from the donor file. However, the upside is that these variables make the CIA more plausible, if they explain association between the EQLS and EWCS variables of interest.

The less matching variables we add to the model, the more accurately we can match the EQLS and EWCS respondents in terms of the matching variables in the model. However, if we do not include variables crucial to the CIA, this assumption becomes less plausible.

Then, for each model specification there are 4 versions:

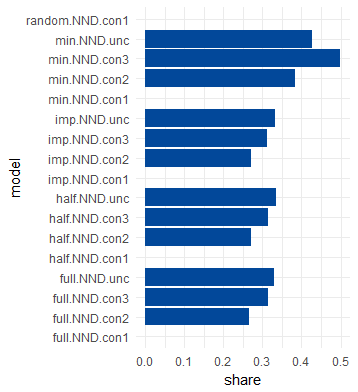
1. Constrained to one donor: donors can only be matched once. This is a strict requirement and makes having donation classes of country, work-life balance and children in the household at the same time impossible because there are not enough donors in each cell. Therefore, children in the household is considered a matching variable in this model and work-life balance is dichotomised to very well and all other categories.
2. Constrained to using the same donor not more than twice. Work-life balance is not dichotomised, but children in the household is a matching variable
3. Constrained to three: the same donor may not be used more than three times and donation classes are not limited in any way
4. Unconstrained: donors can be redrawn unlimitedly; donation classes are not restricted in any way.

For the more constrained versions, we would expect to see that having children in the household does not always have the same value in the EQLS as in the EWCS because we consider it a matching variable. Also for the dichotomised version of work-life balance (in model version a) we may see differences within the four categories of work-life balance. For the other models, the donation classes will make sure that the values of these variables are identical for the EWCS donor to the EQLS recipient.

The distance between the values of the matching variables in the EWCS and in the EQLS can be minimised according to different functions in R-package ‘StatMatch’. For all models, Gower’s dissimilarity index has been used. It is an average of the distances computed on the single variables according to different rules, depending on the type of the variable. This is particularly useful for the matching variables used in these models as they are interval, ordinal and categorical variables. The variable working hours has been standardised in both surveys before being included as a matching variable to correct for the different levels. We are thus assuming that the levels of working hours are not equivalent, but that the distributions are. Finally, the post-stratification weight of the EWCS is used in the matching process. If multiple donors can be considered, donors with a higher post-stratification weight will have a higher probability of being selected.

Figure 11 shows the share of duplicate donors for each model. The constrained versions have no duplicate donors and for the full, important, and half model versions the share of duplicate donors ranges from roughly a quarter to a third depending on the constraint. Most duplicate donors can be found in the minimal model. This is because the distances are based on only two matching variables, which are not differentiating enough to result in a more heterogeneous group of donors.

Figure 11: Share of duplicate donors by model and version



# 7: Evaluation

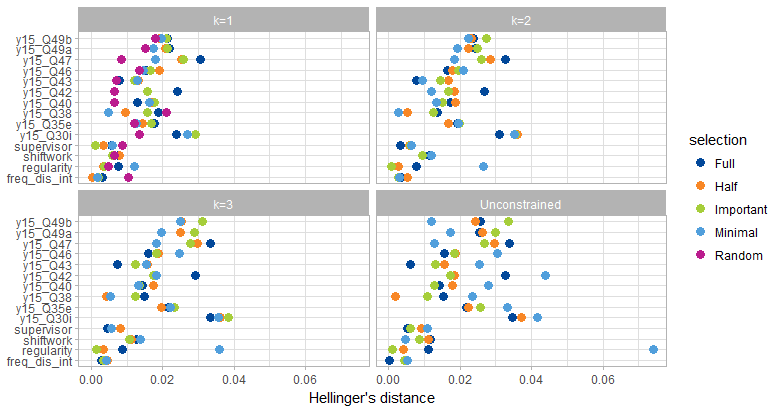
The final step of statistical matching is evaluation and in this step all the models will be evaluated.

## Preserving distributions

The first thing to test is if the distributions of the variables are preserved. Figure 12 shows Hellinger’s distance based on the difference in the distribution of the EWCS variables in the EWCS donor file and the distribution in the synthetic micro data file. This shows that the distributions have been preserved as all variables have values well below the threshold of 0.5, with a few exceptions.

In the upper-left quadrant of the chart, the random model is included. This model shows very low values of Hellinger’s distance and that is because this matching model has no matching variables, it’s simply a random sample of donors in each country weighted by the post-stratification weight and there most capable of reproducing the original distributions.

Between the other models, no clear pattern seems to emerge. However, the less constrained the model in terms of the number of times a donor may be duplicated (as seen in the other quadrants) the bigger Hellinger’s distance. The same reasoning applies here: using the same donors more than once is moving away from simple random sample without replacement.

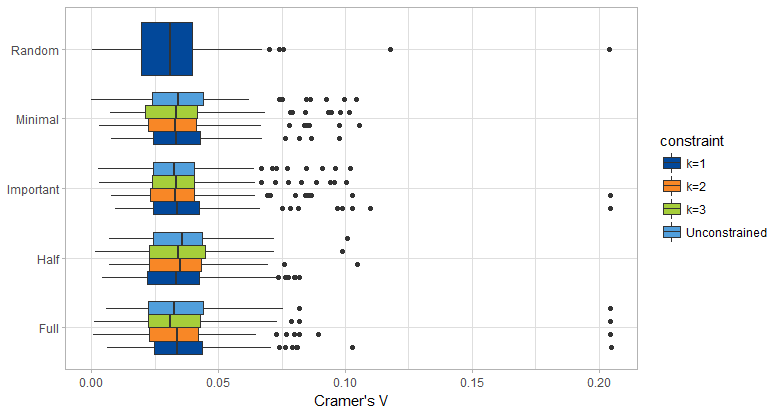
Figure 12: Hellinger’s distance by variable, model and model version

## Association between matched variables

The true joint distribution of the EQLS variables Z and the EWCS variables Y is unknown and the reason for statistical matching in the first place. Therefore, we cannot test if the observed joint distribution in the fused dataset is equal to the true joint distribution. However, we assume that the matching variables X explain the association between the Z and Y variables. Therefore, we would expect to see a difference in the association between the Z and Y variables between the different models based on different sets of matching variables.

We can calculate a measure of association (Cramer’s V) between all combinations of the categorical EWCS variables and the categorical EQLS variables, by each model and each model version. Figure 13 shows the results of this and more specifically the distribution of the Cramer’s Vs by model type and version. Each point in each distribution represents a Cramer’s V of one of the EWCS variables Y with one of the EQLS variables Z.

Figure 13: Distribution of Cramer’s V for each variable combination by type of model and constraint



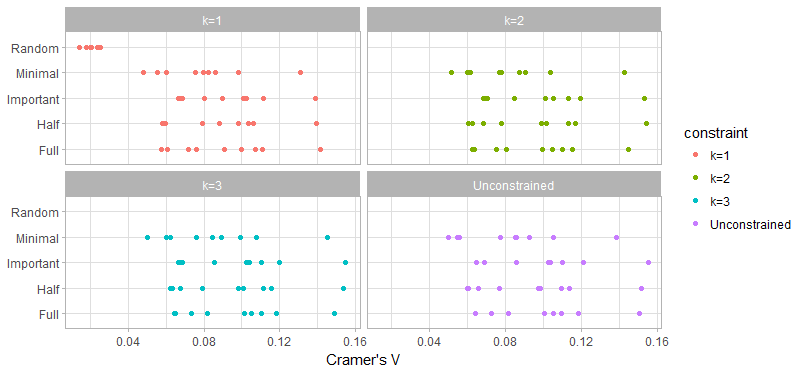
At first glance, a few things stand out. First is that all matching models show higher levels of association than the random model. If in the true distribution there is association between Y and Z, we would also expect the other matching models to show more association if this association can be explained by the matching variables X. Between the matching models other than the random one, however, there is not much difference in the measure of association, indicating that the difference in matching variables may not be very important to the CIA.

The second thing that stands out in this chart is that Cramer’s V is generally low in all models. It is unclear whether this is the result of violating the CIA, or whether the EWCS and EQLS variables are simply not correlated to each other to a large degree. As can be seen by the outliers, certain combinations of variables show more association.

Attempting to show a counterfactual, we can consider matching the some of the variables that were identified as potential matching variables but dropped because they were not considered similar enough. For these variables we would not expect perfect correlations, but we would expect high correlations after matching. Figure 14 shows Cramer’s V for all combinations of the three questions about work-life reconciliation (feeling too tired after the job, work conflicts with family responsibilities, not able to concentrate) that were present in both surveys but with slightly different wording and answer categories. This chart shows stronger association and a bigger difference with the random matching model. The highest scores are for the combinations of the same questions and these show scores in the range of 0.15.

From this we can conclude that the matching model works, but the true correlations between the EQLS and the EWCS variables are small, leading to small differences in Figure 13 compared to the random matching model. Or perhaps the true correlations are bigger, but the matching variables are not adequate for satisfying the conditional independence assumption for these variables and the matching variables satisfy the CIA better when considering the work-life reconciliation questions.

Figure 14: Distribution of Cramers V for each variable combination by type of model and constraint. Variables about work-life reconciliation only.



# Conclusions

EWCS and EQLS share many common variables. This certainly opens the door for statistical matching. However, it is unclear whether the low association between the EWCS and EQLS variables reflects the true distribution or of it is the result of a matching procedure that is not fit for purpose. With regards to the latter case improving the accuracy of the models could help. However, the predictors that can be used for that are limited to the variables that both datasets have in common.

In both scenarios, the added value of statistical matching seems to disappear. This does not mean, however, that statistical matching as a method would not be useful, this simply depends on the datasets and the variables of interest.

# References

Breiman, L. Machine Learning (2001) 45: 5. <https://doi.org/10.1023/A:1010933404324>

European Working Conditions Survey [website](https://www.eurofound.europa.eu/surveys/european-working-conditions-surveys).

European Quality of Life Survey [website](https://www.eurofound.europa.eu/surveys/european-quality-of-life-surveys)

Eurostat (2013), Statistical Matching: A Model Based Approach for Data Integration. Methodologies and Working papers, Eurostat, Luxembourg.

D’Orazio, M. (2015) Statistical Matching and Imputation of Survey Data with StatMatch, R package vignette.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical matching: Theory and practice.* John Wiley & Sons.

R-package randomForest [https://cran.r-project.org/package=randomForesthttps://cran.r-project.org/package=randomForest](https://cran.r-project.org/package=randomForesthttps:/cran.r-project.org/package=randomForest)

R-package StatMatch <https://cran.r-project.org/package=StatMatch>

# Variables

## EWCS variables (Y)

Supervisor: being a supervisor or not

Q26: How many days per week do you usually work in your main paid job? (number)

Q30i: Does your mean paid job involve: Working with computers, laptops, smartphones etc. (7pt scale all of the time to never)

Q35e: how often you have worked in your own home (5pt scale)

atypical: index of atypical working time patterns (night, weekends, etc) (scale 0-100)

regularity: regularity of working hours (same day, same week, etc) (low, medium, high)

Q38: In the last month, has it happened at least once that you had less than 11 hours between the end of one working day and the start of the next working day? (Y/N)

Q40: Over the last 12 months, how often have you been requested to come into work at short notice? (5pt scale)

shiftwork: type of shiftwork (4 categories)

Q42: How are your working time arrangements set? (4 categories)

Q43: Do changes to your working time arrangements occur regularly? (5 categories) (only if fixed schedule)

Q46: Over the last 12 months, how often have you worked in your free time to meet work demands? (5pt scale)

Q47: Would you say that for you arranging to take an hour or two off during working hours to take care of personal or family matters is… (4 pt scale very east to very difficult)

Q49a: does your job involve working at very high speed? (7pt scale)

Q49b: does your job involve working to tight deadlines? (7pt scale)

pace\_det\_3: having more than 3 pace determinants (Y/N)

freq\_dis\_int: having frequent disruptive interruptions or not (Y/N)

Autonomy: scale of autonomy at work (0-100)

## EQLS variables (Z)

Q58d: In general, how would you rate the quality of each of the following public services in your country? Child care services (1-10)

Q58e: In general, how would you rate the quality of each of the following public services in your country? Long term care services (1-10)

Q79: You mentioned that the childcare mainly received by the youngest child is [answer Q78]. How many hours per week is it used? (number of hours) (only for people with children using formal childcare)

childcare: respondents children or children in the household use childcare (only for people with children)

Q78: What is the main type of childcare received by the youngest child (outside of regular school hours)? (five types of childminding)

Q82: To what extent did cost make it difficult for you to use childcare services? (Very difficult, a little diffficult, not difficult at all) (only for people with children using formal childcare)

Q68a: Have you, or someone close to you, used the following services in the last 12 months? Here we are asking about formal services, not care provided by families. Nursing care services at your/this person’s home (yes I have, yes someone close to me has, nobody has)

Q68b: Have you, or someone close to you, used the following services in the last 12 months? Here we are asking about formal services, not care provided by families. Home help or personal care services in your/this person’s home (yes I have, yes someone close to me has, nobody has)

Q68c: Have you, or someone close to you, used the following services in the last 12 months? Here we are asking about formal services, not care provided by families. Residential care or nursing home (yes I have, yes someone close to me has, nobody has)